基于稀疏分布式表征的英文著者姓名消歧研究 *

翟晓瑞，韩红旗†，张运良，李 仲

(中国科学技术信息研究所 富媒体数字出版内容组织与知识服务重点实验室，北京 100038)

摘 要：为将稀疏分布式表征理论应用到著者姓名消歧，了解其在解决姓名消歧问题时的效果，提出了基于稀疏分布式表征的英文文献著者姓名消歧方法。该方法选择论文摘要文本信息作为消歧特征，将其生成二进制表示的 SDR 码。根据待消歧论文的 SDR 与同名作者的论文 SDR 相似度对比来实现著者姓名消歧。最终得到的结果为准确率 98.21%，召回率 76.75%，F 值 86.17%，证明提出的消歧方法具有较好的效果。通过对比该方法与利用合著者特征进行消歧的方法，说明该方法能够较好地解决文献著者姓名歧义问题。此外，该方法还可将作者未收录在作者库中的论文识别出来并将其指派给新作者，无须重新学习和更新模型。

关键词：姓名消歧；稀疏分布式表征；语义指纹；层级时序记忆模型

中图分类号：TP391.1 doi: 10.3969/j.issn.1001-3695.2018.07.0380

Research on English author name disambiguation based on sparse distributed representation

Zhai Xiaorui, Han Hongqi, Zhang Yunliang, Li Zhong

(*Key Laboratory of Rich-media Knowledge Organization & Service of Digital Publishing Content*, *Institute of Scientific & Technical Information of China*, *Beijing* 100038, *China*)

Abstract: In order to apply the Sparse Distributed Representation theory to the author name disambiguation, and to know the effect of the theory in solving the name disambiguation problem, this paper proposed a method based on Sparse Distributed Representation to disambiguate English author name. This paper selected summary as disambiguation feature and generated binary representation of SDRs. And then it constructed the similarity matrix based on the similarity comparison of the training set, the experiment is performed after the appropriate threshold set. The final accuracy is 98.21%, the recall is 76.75%, and the F-value is 86.17%. The result indicates that the proposed method has a good effect. By comparing the method proposed with the method based on co-authors, it can be concluded that the method proposed can better solve the ambiguity problem of author names. In addition, the method can also identify the papers whose authors are not included in the author database, and assign to new authors without relearning and updating the model.

Key words: name disambiguation; sparse distributed representation; semantic fingerprint; hierarchical temporal memory

0 引言

由于现实中不同人物具有相同的姓名、同一人物存在多种姓名变体等原因，姓名歧义现象非常严重，影响到了人们有效地获取和利用相关信息[1]。随着网络数据的快速增长，依靠手工解决姓名歧义问题变得不可能，因此，如何利用自然语言处理、机器学习等自动化技术消除姓名歧义是研究人员必须面对的重要挑战，该问题也成为近年来的研究热点之一[1]。为了推动自动化姓名消歧技术的发展，国内外相关机构组织了专门的竞赛评测会议，取得了一定的影响，推动了姓名消歧问题的的解决，如网页人物搜索评测竞赛(Web People Search Evaluation Campaign,WePS)、CLP2010(Chinese Language Processing 2010 年)[1]等。

文献著者姓名歧义是姓名歧义的一种，主要表现在当检索文献数据库[2]、机构知识库[3]等某位作者的科研成果时，系统会将所有同名作者的科研成果全部返回，降低了检索准确度[4]，给信息用户利用信息带来了困扰。文献著者姓名消歧旨在消除这种姓名的歧义性，即判断同名作者下的各文献归属于现实中的哪一个人物实体的处理过程，是利用文献进行合著者社会网络构建、科研能力评估、学术推荐等研究一个必不可少的环节

收稿日期：2018-07-18；**修回日期：**2018-09-12 **基金项目：**国家自然科学基金资助项目（71473237）；中国工程科技知识中心建设项目（CKCEST-2018-1-26）

作者简介：翟晓瑞（1993-），女，河南灵宝人，硕士研究生，主要研究方向为文本挖掘；韩红旗（1971-），男（通信作者），副研究员，博士，主要研究方向为文本挖掘与社会网络分析、知识组织与知识工程（bithhq@163.com）；张运良（1979-），男，研究员，博士，主要研究方向为知识组织、知识服务、文本分类；李仲（1985-），男，硕士研究生，主要研究方向为文本挖掘与社会网络分析.

[2,5]。与网页姓名消歧不同，文献著者姓名消歧要提供更多的人物信息和某一人物个体全面的特征，这超出了要指出网页或文章中是否提及某个人物的范畴，以及分类、聚类的任务要求[6]。

现有的著者姓名消歧方法大多利用题目、关键词等短文本信息作为作者的研究方向[5]或论文的主题信息[7]进行消歧，对于摘要等长文本信息却未能加以利用，相比关键词和题目等短文本信息而言，摘要更能完整的表达一篇论文的核心思想、所用方法等信息，是作者研究主题的集中体现。虽然有些聚类方法虽然选择了摘要作为消歧特征之一，但经过分词、去除停用词、特征词抽取等操作将长文本信息又转变成了词的形式[4, 8, 9]，一定程度上损失了语义信息。此外现有的消歧方法主要利用聚类算法作为划分方法[10]，而有些聚类方法并不能确定同名的人数[11]，因而在选择初始聚类个数时难度较大，而最初聚类的划分将直接影响最终聚类的效果[12]。现有消歧方法大多基于已有的数据集，通过分类、聚类进行消歧，而对于新收录的论文无法利用之前已经产生的分类信息直接进行区分，需要重新进行学习，不断更新模型，消歧效率较低，花费时间较长[2, 5]。

稀疏分布式表征（sparse distributed representation,SDR）理论是生物神经网络层级时序记忆（hierarchical temporal memory,HTM）理论[13]的主要信息表示方式，它可将文本信息转换为承载内容语义特征、具有某固定长度（如 16 384 位）的二进制序列，序列中的“1”代表活跃位，所占比例一般为 0.05%~2%，它将文本信息之间的相似性比较转换为二进制序列形式的语义指纹的相似性度量。SDR 因其较高的位数和较低的稀疏度，具有较高的鲁棒性以及较低的误配率，可有效用于不同文本内容的匹配计算中[14]。利用 SDR 的这种特性和论文摘要对作者研究主题的代表性，可以对同一作者的研究成果进行辨识，从而实现对著者姓名的消歧。因此，本文提出了基于 SDR 的英文著者姓名消歧方法，旨在探索该理论在解决姓名歧义问题时的可用性及消歧效果。

1　国内外研究现状

1.1　姓名消歧研究现状

自 1998 年 Bagga 和 Baldwin[15]首次提出跨文档姓名消歧以来，姓名消歧已成为当今国内外学者研究热点之一。通过在万方数据库和 Web of Science 数据库的文献调研，本文将近年来常用的文献著者姓名消歧方法主要分为两大类，即基于特征的姓名消歧方法和基于机器学习的姓名消歧方法。

基于特征的方法，主要是通过选择作者的个人信息或是论文的题录信息作为消歧特征，将文献之间的相似性比较转换为消歧特征之间的比较，从而判断文献的归属。其中，常用的作者个人信息包括机构、电子邮件、联系方式等，论文的题录信息包括题目、关键词、合著者、研究方向等[7]。Singh[16]从专利数据中抽取了发明人姓和地区两个字段，利用 if-else 判定规则和字符串精确匹配来判断两条著者记录对是否正确匹配。Fleming 等人[17]抽取专利权人和发明人地区字段并将其合并，利用 if-then-else 匹配规则和字符串精确匹配，结合设定的阈值，判断两条著者记录对是否正确匹配。线岩团等人[18]选择专有名词、姓名、机构、地名、题目、职业、职称以及上下文中以名词词组形式出现的概念作为消歧特征，通过构建相似度矩阵，采用近邻传播聚类算法进行消歧。张雄等人[7]选择文献合著者信息、姓名关联信息和主题信息作为消歧特征，采用多特征融合的方法实现姓名消歧。李孟亚[19]从图书作者简介信息中抽取出实体特征、上下文特征和社会关系特征等三类特征，使用属性互斥放大和特征空缺缩小方法计算作者相似度，最后通过凝聚层次聚类完成消歧。

基于机器学习的方法又可分为三类，即监督式学习、无监督学习和半监督学习。基于监督学习的方法通过训练标签数据集训练分类器，继而进行姓名消歧，Kim 等人[20]采用随机森林和 DBSCAN 聚类的方法，在 USPTO（United States Patent and Trademark Office，美国专利商标局）专利发明人姓名消歧竞赛的数据集上进行实验，得到的结果优于竞赛结果。基于无监督学习的方法是根据相似度计算方法在无标签训练集中进行聚类，将相似度比对结果满足阈值要求的视作同一作者，朱亮亮[8]利用改进的 K-means 算法，根据最大最小原则初始聚类中心，克服了传统 K-means 算法随机选择初始聚类中心一定概率导致局部收敛的问题。基于半监督学习的方法通常采用小数据量标签数据和大数据量无标签数据来训练模型，继而进行消歧，Ronald 等人[21]借助于 Torvik 和 Smalhesie[22]的方法，通过统计产生准确度较高的人造标签数据集，并在贝叶斯框架下使用逻辑回归方法判断作者记录对的匹配情况。

1.2　SDR 相关理论

语义指纹技术可将文本内容和特征映射为固定长度的二进制数字字符串，一般为 32 bit、64 bit 和 128 bit，以此来表示文本的特征，将文本间的相似性比较转换为二进制序列的距离度量，常用在网页去重、数据文档抄袭检测等应用中[23]。现有的语义指纹算法生成的二进制序列多为 32 bit、64 bit 和 128 bit，对于存储空间有较少的要求，且具有较低的鲁棒性。

SDR 是 Numenta[1]公司提出的 HTM 理论的关键组成部分[24]，是语义指纹的一种形式。SDR 的生成算法基于语义折叠理论[14]，实现了文本到语义指纹的转换。它利用构建的语料库获取单词的上下文文本片段，并将其映射在一个二维矩阵中，使得主题相似的文本片段在矩阵的位置较近，主题不同的在矩阵中的位置较远。然后将该矩阵展开为一维向量，就可以生成单词的 SDR 代码了。具体来说，对于一个单词，若出现在对应的文本片段中，则 SDR 向量对应的位置为 1，否则为 0。该向量基于上下文表示单词的语义含义，然后组合各个单词向量以形成句子向量，还可形成文本向量[14]。SDR 向量是多维向量，其

[1] http://www.numenta.com

长度 n 一般在 1024~65536，其中“1”的位数ω在 10 至 40 位，即控制稀疏度在 0.05%至 2%之间。SDR 的每一位都有一定的语义意义，如果两个 SDR 在同一位置均为 1，则说明这两个 SDR 共同拥有该位对应的属性[25]。

SDR 在存储时，仅存储活跃位的信息即可，大大减少了对存储空间的需求。SDR 理论中使用重叠数（overlap）来定义两个 SDR 编码的相似性，即是指两个向量在相同位置都是“1”的个数。当重叠数超过某个阈值θ时，则认为这两个 SDR 时匹配（matching）的。当 n=1024，ω=9 时，两个 SDR 向量的错误匹配概率就已经降到了 3.0365*10-22，所以说 SDR 编码的鲁棒性很高，并且具有一定的容错能力，即使丢弃或移动了一些位，其代表的语义也会保持不变[14]。

为了方便研究人员使用 SDR，Cortical.io 公司提供了名为 Retina 的 API，实现了基于 SDR 理论的语义指纹生成，它将输入的文本信息通过语义折叠方法得到其对应的 128*128 维矩阵，再将其表示为向量的形式，即 16 384 位的 SDR 码，然后将值为“1”的位对应的索引下标作为返回值提供给用户，用户可依此来生成对应的 SDR 码。

2　基于 SDR 的英文著者姓名消歧方法

本文提出的基于 SDR 的英文著者姓名消歧方法，通过比较一篇待消歧论文与已消歧论文之间的相似性来判断该待消歧文献是哪一位同名作者的论文，从而将待消歧论文与相应的人物实体相对应，实现将同名作者的论文彼此区分开来的目的。提出的方法由 SDR 生成、SDR 比较、作者匹配、争议仲裁、作者指派等五个处理过程组成，如图 1 所示。

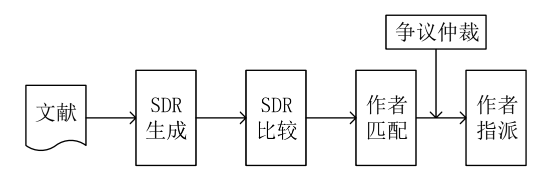


图 1　基于 SDR 的英文著者姓名消歧方法流程

2.1　特征选择及 SDR 生成

文献数据库中检索到的文献题录信息一般包括题名、作者、合著者、作者机构、期刊名、摘要、关键词等信息[26]，其中摘要为文本信息，且高度概括了文章内容，一定程度上代表了作者的思想，是表征作者信息的重要特征，故本实验选择题录信息中的摘要信息作为消歧所用的特征。在获得摘要信息后，利用 SDR 生成算法将摘要文本信息生成 SDR 码，作为消歧使用的语义指纹。具体过程如图 2 所示。

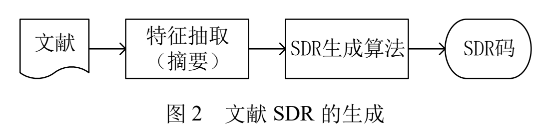


图 2　文献 SDR 的生成

2.2　SDR 比较

在获取一篇论文摘要信息且生成其 SDR 后，本文将其 SDR 与现有同名作者的全部论文的 SDR 进行相似度比较，两个 SDR 相似度的计算采用 cortical.io 提供的方法[27]，比较结果记为 H(x)。H(x)的值为介于 0~1 的小数，SDR 取 0 时表示两个 SDR 对应的论文确定不属于同一作者，为 1 时表示两个 SDR 对应的论文确定属于同一作者。

2.3　匹配方案

对于一篇待消歧的论文 p，与已消歧的一个同名作者α的 N 篇论文的 SDR 进行比较，得到 N 个相似性比较结果，即 Hi(x)（i=1，2，…，N）。为了确定 p 是否为作者α的论文，设置了相似性比较阈值区间(δ_1, δ_2)，δ_1 和 δ_2 的取值需要根据实际情况确定。确定论文 p 是否为一个作者α的论文的过程如图 3 所示，详述如下：

a）当待消歧论文 p 与已消歧作者α的第 i 篇论文 SDR 的比较结果 $H_i(x)$大于阈值 δ_2 时，确定论文 p 的作者是α。虽然 p 与作者α的 N 篇论文存在 N 次 SDR 比较，但这种情况只要出现一次，就可认定论文 p 的作者是α。

b）当全部 N 次 SDR 比较的相似度 Hi(x)小于阈值 δ_2 时，这时存在两种情况，一种是存在 $H_i(x)$，使得 $\delta_1<H_i(x)<\delta_2$，这种情况下认为论文 p 的作者可能是α，另一种是任一 $H_i(x)$均小于 δ1，这种情况下认为论文 p 的作者不可能是α。统计 $H_i(x)$的值位于区间(δ_1, δ_2)的情况出现的数量，即计算 $H_i(x)$在(δ_1, δ_2)内的个数，记为 n，若 n/N>h，则文献 p 的作者为α。h 是一个阈值参数，需要根据实际情况确定。

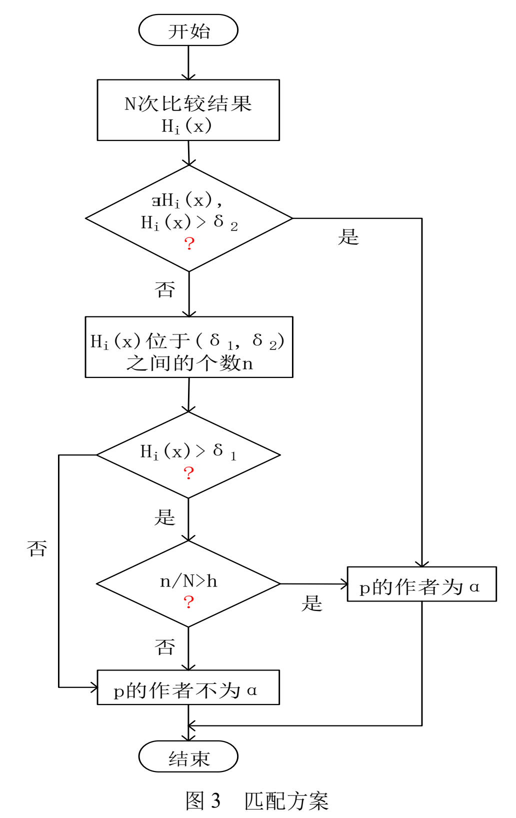


图 3　匹配方案

2.4　指派方案

论文 p 在经过匹配后，设其与 m 位作者相匹配，则存在如下三种可能结果：

a）m=0，即文献 p 未能与已有作者匹配，则将文献 p 指派给一名新作者；

b）m=1，即文献 p 只与一位作者匹配，则将文献 p 指派给

该作者；

c）m>1，即文献 p 同时与多位作者匹配，此时，可由仲裁程序判定文献 p 指派给哪一位作者。

不失一般性，假设文献 p 同时与作者 α_1 和作者 α_2 匹配，分别计算文献 p 与两位作者所有文献的相似性比较结果的平均值，若 $\frac{\sum H(\alpha_1)}{N_{\alpha_1}} > \frac{\sum H(\alpha_2)}{N_{\alpha_2}}$，则将该文献指派给作者 α_1，否则就指派给作者 α_2（见图 4）。对于存在多于两个以上的指派，则取相似性比较结果的平均值最大的作者作为指派对象。

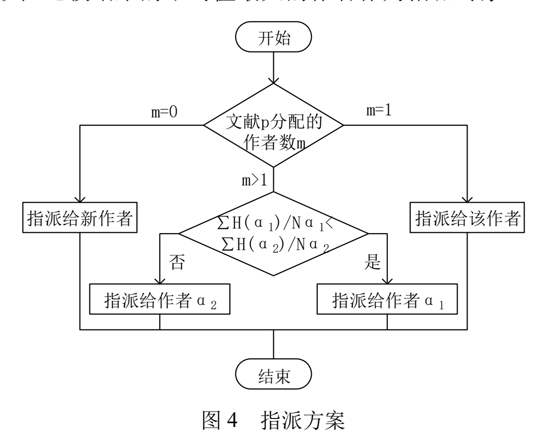


图 4　指派方案

2.5　评价指标

为评价基于 SDR 的姓名消歧方法的有效性，本文采用了由中国中文信息学会[2]（CIPS，Chinese Information Processing Society of China）与国际计算语言学协会中文处理专业兴趣组[3]（SIGHAN）于 2012 年主办的中文处理国际会议（CLP-2012）中使用的准确率（precision）、召回率（recall）及 *F* 值（*F*-measure）评价指标。

其中，准确率是指识别出归属为作者 α 的文献中实际为作者 α 的文献所占的比率，召回率是指实际为作者 α 的文献中识别出来的文献所占的比率。两者的取值在 0 到 1 之间，越接近 1，效果越好。但由于两者在实际中常常是相互影响的，提高一个指标会带来另一个指标的降低，因此需要采用 F 值来综合反映整体的指标。

如果把姓名消歧结果看做是簇，每一个簇是同一个作者的结果集合，则每一个簇的准确率和召回率计算公式如下：

$$Precision_i = \frac{|S_i \cap R_i|}{|S_i|}$$

$$Recall_i = \frac{|R_i \cap S_i|}{|R_i|}$$

其中:R 表示人工消歧的结果集合，$R_i \in R$ 表示人工消歧的结果集合中的某一簇。S 表示利用消歧方法消歧的结果集合，$S_i \in S$ 表示利用消歧方法消歧的结果集合中的某一簇。两个集合的大小分别表示为$|R_i|$、$|S_i|$。

在得到每一簇的 $precision_i$ 和 $recall_i$ 后，将他们的平均值作为整体消歧效果的准确率和召回率，其中 N 代表簇数。

$$precision = \frac{1}{N}\sum_{i=1}^{N} Precison_i$$

$$recall = \frac{1}{N}\sum_{i=1}^{N} Recall_i$$

整体消歧效果的 F 值的公式如下：

$$F-measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

3　基于 SDR 的英文著者姓名消歧实验

3.1　数据集构建

为了验证提出的方法的效果，利用 ResearchGate 学术社交网络、通过手工方式构建了实验数据集。选择了四个具有较高歧义性的名字进行数据集的构建，分别为 J Huang、L Stevens、T Joe、J. Baker，其中一个为中文著者的英文姓名，三个为外国人姓名。在对获取的作者论文信息进行初步分析后，发现其中部分文献数据缺少摘要信息，所以将这些数据从数据集中剔除，最后的数据集包含 19 位作者的 88 篇文献。将数据集分为两部分，分别归入数据集 1 和数据集 2，其中数据集 1 中包含 17 位作者 47 篇文献，数据集 2 中包含 18 位作者 41 篇文献。再从数据集 1 和数据集 2 分别抽取部分数据组建训练集，标记为训练集 1（D1）与训练集 2（D2），D1 中共计 7 位作者 23 篇文献，D2 中共计 7 位作者 19 篇文献。训练集 1 和训练集 2 被用来估计阈值，即 δ_1、δ 和 h。数据集 1 和数据集 2 被用来进行实验，以测试本文提出的基于 SDR 的姓名消歧方法的效果。

为了实验上的方便，利用 Cortical.io 公司的 Rentina API 为每一篇论文的摘要文本生成了其 SDR 的表示形式。Rentina API 无须对文本信息进行分词及停用词处理，但实际文献中存在的一些特殊字符，或者部分非英语国家的作者在撰写论文时引入的非英文状态下的符号会对 SDR 结果产生影响，因此还需要对摘要文本进行适当的规范化处理。图 5 是左边为 Rentina API 返回的数字序列，每个数字代表了 SDR 向量中值为“1”的索引下标。图 6 是依此生成的实际 SDR 码。

3.2　阈值选择

将 D1 与 D2 中的同名作者的文献 SDR 码进行一一对比，得到相似性结果矩阵。比较分析同一个作者的任意两篇文献的相似性比较结果，与不同作者的任意两篇的相似性比较结果。如图 7 所示。粗线框内的为同一作者的两篇文献之间的相似性比较结果，剩下的为不同作者的任意两篇文献的相似性比较结果。

分析结果可知，归属于同一作者的任意两篇文献的比较结果在(0.1522，0.5833)间，主要集中在(0.42，0.52)间；归属于不同作者的任意两篇文献的比较结果在(0.1657,0.4919)间，主要集中在(0.29,0.44)间,如图 8 所示。

因此，将阈值的选择区间设定在(0.42，0.52)间，δ_1 取 0.42,

[2] http://www.cipsc.org.cn/

[3] http://sighan.cs.uchicago.edu

为了判断最佳的参数δ和 h，对相似性比较结果矩阵进一步分析，使δ2 取值从 0.42 开始，以 0.01 为步长增加，同时 h 分别取 20%、30%、40%，观察不同组合下得到的查全率、召回率及 F 值曲线，找到消歧效果最好的参数组合。曲线图如图 9 所示，最佳消歧效果对应的阈值组合为δ_1为 0.42，δ_2为 0.50，h 为 20%。

```
[16, 18, 19, 27, 30, 34, 61, 70, 71, 75, 76, 77,
458, 551, 557, 563, 564, 641, 687, 703, 722,
1332, 1350, 1397, 1400, 1402, 1409, 1438,
2031, 2059, 2060, 2116, 2118, 2156, 2158,
2640, 2678, 2706, 2719, 2782, 2784, 2818,
3325, 3326, 3372, 3420, 3432, 3437, 3451,
3976, 3978, 4068, 4091, 4092, 4094, 4095,
4478, 4479, 4489, 4513, 4578, 4624, 4626,
5421, 5471, 5472, 5478, 5482, 5487, 5523,
6777, 6848, 6905, 7233, 7338, 7427, 7465,
```

图 5　API 返回的数字序列

```
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0,
0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,
0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1,
0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0]
```

图 6　生成的 SDR 样例

分组序号	2-1	2-2	2-3	2-4	2-11	2-12	2-13	2-14
1-1	0.5427	0.5163	0.5325	0.5457	0.4736	0.4746	0.3679	0.3028
1-2	0.5356	0.4949	0.5030	0.5122	0.4187	0.4451	0.3760	0.2846
1-3	0.4959	0.4797	0.5691	0.4929	0.4299	0.4238	0.3679	0.3100
1-4	0.5112	0.5213	0.4766	0.5203	0.4319	0.4268	0.3526	0.2764
1-11	0.4898	0.4390	0.4583	0.4116	0.5549	0.4980	0.2947	0.2297
1-12	0.4685	0.4339	0.4228	0.4085	0.5061	0.4675	0.3110	0.2419
1-13	0.4360	0.4309	0.4177	0.3923	0.5508	0.4878	0.2907	0.2073
1-14	0.4228	0.4004	0.4614	0.4126	0.3923	0.3933	0.4217	0.3648
1-15	0.4157	0.4045	0.4197	0.4004	0.3943	0.3720	0.3943	0.3272
1-16	0.4157	0.4024	0.4258	0.4065	0.3994	0.3506	0.4421	0.4075

图 7　相似性比较结果矩阵样例

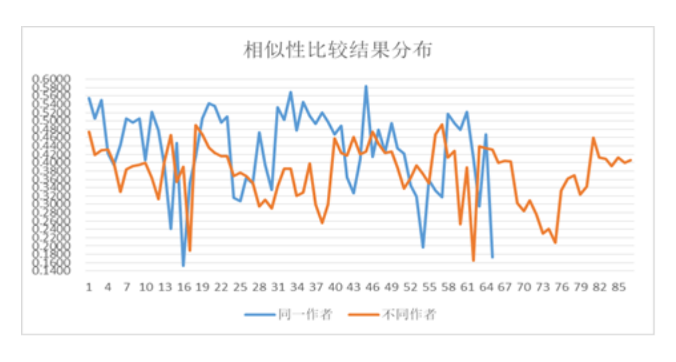


图 8　相似性比较结果分布

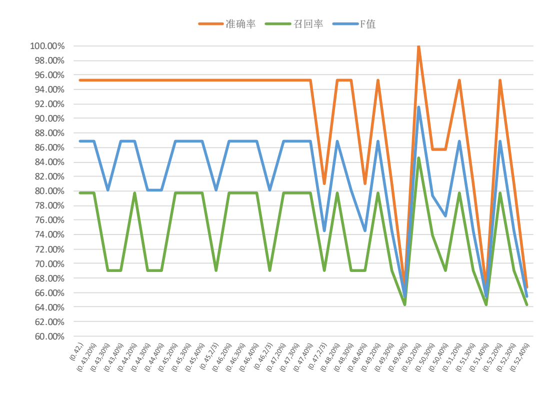


图 9　不同阈值组合消歧结果

3.3　基于 SDR 的消歧实验结果

根据提出的实验方案，将数据集 1 和 2 中同名作者的任意两篇文献 SDR 码进行比较，得到相似性比较结果矩阵，再利用得到的阈值，即δ_1=0.42，δ_2=0.50，h=20%，进行姓名消歧实验。

若 H(x)>0.50,则将该文献分配给该作者；若 0.42<H(x)<0.50，则计算 H(x)位于(0.42，0.50)内的百分比，若高于 20%，则与对应作者相匹配。若仅与一名作者相匹配，则将该文献指派给作者；若同时与多名作者相匹配，则比较其 H(x)的平均值，将文献最终指派给平均值高的对应的作者；若未能与已有作者相匹配，则将该文献指派给新作者。

实验最终得到的准确率为 98.21%，召回率为 76.75%，F 值为 86.17%。

3.4　与基于合著者特征的消歧实验对比

在文献的题录数据中，除了本文采用的摘要文本信息外，还有一些特征也被广泛地用于姓名消歧实验中，如合著者特征、作者机构等。根据张雄等人的研究[7]，合著者特征消歧达到了较好的效果。故选择了合著者特征进行消歧实验，将其结果作为对比，来评价基于 SDR 的姓名消歧效果。

基于合著者特征的姓名消歧实验中，首先人工对合著者姓名进行规范化处理，避免合著者姓名的歧义问题对实验结果造成影响，然后利用字符串匹配的方法进行消歧。若两篇同名作者的论文中至少存在同一个合著者，则认为这两篇文献的作者为同一人。

经过实验，得到的准确率为 98.32%，召回率为 73.68%，F 值为 84.24%。

图 10 展示了本文提出的方法与合著者特征方法的对比，两者在准确率上差别不大，合著者特征高一点，但召回率上本文方法有较明显的优势，从而在 F 值对比上也取得较好的优势。本实验所用数据集中独著论文数量较少，使得合著者特征方法取得了较好的准确率，如果存在较大比例的独著论文，则合著者特征方法的准确率可能会大幅下降。

对于作者未收录在数据集 1 中的论文，合著者特征方法无法对其进行消歧，而本文提出的方法可将其识别出来并指派给新作者。但对于存在合著关系的同名作者，两种消歧方法均失效。

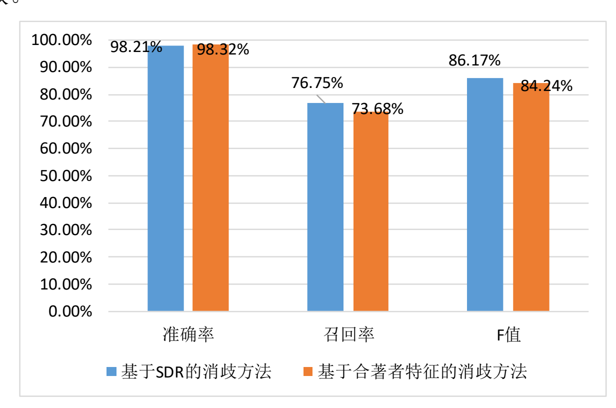


图 10　不同消歧方法的消歧效果柱状图

4　结束语

文献著者姓名消歧是科研成果评价、合著者社会网络构建、知识服务系统构建等问题的基础性研究。本文提出的基于稀疏分布式表征的英文文献著者姓名消歧方法，选择摘要信息作为消歧特征，利用 SDR 生成算法将其生成 16 384 位 SDR 码，通过 SDR 码的相似性比较得到比较结果，在确定阈值参数后，将满足条件的论文指派给相应的作者。最终得到的实验结果为准确率 98.21%，召回率 76.75%，F 值 86.17%，证明稀疏分布式表征可有效用于姓名消歧。本文提出的方法可有效识别出作者未收录在作者库中的论文，并将其指派给新作者。

虽然本研究在构建的实验数据集上取得了较好的效果，初步验证了基于 SDR 理论进行著者姓名消歧的可行性和有效性，但仍然存在一些不足之处。存在的主要问题有两个，第一个是实验使用的数据集规模小，无法涵盖实际中文献著者的多种情况，可能缺乏全面性及代表性；第二个是消歧过程中将部分作者指派为新作者，但未将新发现作者同步更新至已消歧数据库中，虽然在本文的实验这种不及时更新未造成大的影响，但在大规模数据实验中可能出现归类为新作者中的部分作者实际中为同一人物实体的情况，从而对消歧效果产生明显的影响。

本文采用的方法是付媛论文[23]方法的改进版本，总体思路上相同，主要区别如下：a）本文方法与付媛方法中语义指纹的生成方式不同。付媛的方法中，选择哈希函数将论文文本中的词汇生成哈希值，通过 Simhash 算法生成论文文本的语义指纹；本文方法中，词汇的 SDR 是基于大规模语料学习生成的，论文文本的 SDR 指纹基于词汇的 SDR 生成；b）指纹的匹配方案不同。付媛论文实验中设置了一个 δ 和 h 作为阈值，将指纹比较结果中大于 δ 的结果所占比例超过 h 的认定为同一作者；本文方法设置了 δ_1、δ_2 和 h 三个阈值参数，若比较结果大于 δ_2 则认定其为同一作者，若比较结果在 δ_1 和 δ_2 之间，则比较所占比例，超过 h 的认定为同一作者；c）姓名消歧的语言环境不同，付媛的论文对中文著者姓名进行消歧，本文对英文著者姓名进行消歧。

在未来的研究工作中，将对以下几个方面进行深入研究：a)将合著者特征、机构特征等和 SDR 融合消歧，以提升该方法在姓名消歧上的效果，促进其在实际系统中的应用；b)在阈值选择过程中，考虑应用深度学习等算法，优化参数设置，以达到更为理想的消歧效果；c)目前关于 SDR 生成方法的核心代码尚未公布，所以只能通过 Numenta 战略合作伙伴——Cortical.io 公司提供的名为 Retina 的 API，获取英文文本的 SDR 表示。今后将构建中文语料库，研究基于中文语料库的中文词汇 SDR 生成，以实现基于中文文献的著者姓名消歧。

参考文献：

[1] 王鑫. 人名消歧关键技术研究与实现 [D]. 哈尔滨: 哈尔滨工业大学, 2012. (Wang Xin. Research and implementation of key techniques of personal name disambiguation [D]. Harbin: Harbin Institute of Technology, 2012.)

[2] 付媛, 朱礼军, 韩红旗. 姓名消歧方法研究进展 [J]. 情报工程, 2016 (1): 53 - 58. (Fu Yuan, Zhu Lijun, Han Hongqi. A survey of name disambiguation [J]. Technology Inteligence Engineering, 2016 (1): 53-58.)

[3] Piotr A, Szymon S. Person name disambiguation for building university knowledge base [C]// Proc of Asian Conference on Intelligent Information and Database Systems. Berlin: Springer, 2016: 270-279.

[4] 任景华. 利用优化的 DBSCAN 算法进行文献著者人名消歧 [J]. 图书馆理论与实践, 2014 (12): 61 - 65. (Ren Jinghua. Document author name disambiguation by using optimized DBSCAN algorithm [J]. Library Thory and Practice, 2014 (12): 61-65.)

[5] 袁军鹏, 俞征鹿, 苏成, 等. 作者重名辨识研究进展 [J]. 数字图书馆论坛, 2011 (10): 60 - 65. (Yuan Junpeng, Yu Zhenglu, Su Cheng, et all. A survey of author name disambiguation [J]. Library Theory and Practice, 2011 (10): 60-65.)

[6] Neil R S, Vetle I T. Author name disambiguation [J]. Annual Review of Information Science & Technology, 2015, 43 (1): 1-43.

[7] 张雄, 陈福才, 黄瑞阳. 基于融合特征相似度的实体消歧方法研究 [J]. 计算机应用研究, 2017, 34 (2): 347 - 350. (Zhang Xiong, Chen Fucai, Huang Ruiyang. Research on entity disambiguation method based on fusion feature similarity. [J]. Application Research of Computers, 2017, 34 (2): 347-350)

[8] 朱亮亮. 利用改进的 K-means 算法实现文献著者人名消歧 [J]. 软件导刊, 2013 (5): 63 - 66. (Zhu Liangliang. Research on name disambiguation based an improved K-means algorithm [J]. Software Guide, 2013 (5): 63-66.)

[9] 杜婧君. 基于中文维基百科的命名实体消歧方法研究 [D]. 杭州: 杭州电子科技大学, 2012. (Du Jingjun. Research on method of Chinese named entity disambiguation based on Chinese Wikipeida [D]. Hangzhou: Hangzhou Dianzi University, 2012.)

[10] 阳怡林, 陈刚, 周杰, 等. 人名消歧研究综述 [J]. 信息工程大学学报, 2016 (4): 478 - 483. (Yang Yilin, Chen Gang, Zhou Jie, et all. Research on name disambiguation: A survey [J]. Journal of Information Rngineering University, 2016 (4): 478-483.)

[11] 陈晨, 王厚峰. 基于社会网络的跨文本同名消歧 [J]. 中文信息学报, 2011, 25 (5): 75 - 82. (Chen Chen, Wang Houfeng. Social network based cross-document personal name disambiguation [J]. Journal of Chinese Information Processing, 2011, 25 (5): 75-82.)

[12] 阳怡林, 周杰, 李弼程. 基于聚类集成的人名消歧算法 [J]. 计算机应用研究, 2016, 33 (9): 2716 - 2720. (Yang Yilin, Zhou Jie, Li Bicheng. Name disambiguation algorithm based on ensemble [J]. Application Research of Computers, 2016, 33 (9): 2716-2720.)

[13] Ahmad S, Hawkins J. Properties of sparse distributed representations and their application to hierarchical temporal memory [J]. Eprint Arxiv, 2015.

[14] De Sousa-Webber F. Semantic folding theory and its application in semantic

fingerprinting [J]. Computer Science, 2015.

[15] Amit B, Breck B. Entity-based cross-document coreferencing using the Vector Space Model [C]// Proc of Meeting of the Association for Computational Linguistics and, International Conference on Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 1998: 79-85.

[16] Jasjit S. Collaborative networks as determinants of knowledge dDiffusion patterns [J]. Management Science, 2005, 51 (5): 756-770.

[17] King III C, Fleming L, Juda A. Small worlds and regional innovative advantage [J]. Organization Science, 2007 (6): 938–954.

[18] 线岩团，余正涛，洪旭东，等．基于特征加权重叠度的中文实体协同消歧方法 [J]. 中文信息学报，2017, 31 (2): 36－41. (Xian Yantuan, Yu Zhengtao, Hong Xudong, et all. Collaborative entity disambiguation method based on weighted feature overlap relatedness for Chinese [J]. Journal of Chinese Information Processing, 2017, 31 (2): 36–41.)

[19] 李孟亚．基于融合特征的中文图书作者人名消歧方法研究 [J]. 电脑知识与技术，2018 (11): 182－184. (Li Mengya. Research on Chinese book author's name disambiguation based on fusion features [J]. Computer Knowledge and Technology, 2018 (11): 182-184.)

[20] Kim K, Khabsa M, Giles C L. Random forest DBSCAN for USPTO inventor name disambiguation [EB/OL]. (2017-09-14) . https://arxiv. org/abs/1602. 01792.

[21] Li Guancheng, Lai R, D'Amour A, *et al.* Disambiguation and co-authorship networks of the U. S. patent inventor database (1975-2010) [J]. Research Policy, 2014, 43 (6): 941–955.

[22] Torvik V I, Smalheiser N R. Author name disambiguation in MEDLINE [J]. ACM Trans on Knowledge Discovery from Data, 2009, 3 (3): 1–29.

[23] 付媛．基于语义指纹的中文文献著者姓名消歧方法研究 [D]. 北京：中国科学技术信息研究所，2016. (Fu Yuan. Research on Chinese authors name disambiguation based on semantic fingerprint [D]. Beijing: Institute of Scientific and Technical Information of China, 2016.)

[24] Cui Yuwei, Ahmad S, Hawkins J. The HTM spatial pooler-A neocortical algorithm for online sparse distributed coding [J]. Frontiers in Computational Neuroscience, 2017, 11: 111.

[25] Hawkins J, Ahmad S, Purdy S, *et al.* Biological and machine intelligence (BAMI) [EB/OL]. (2018-03-08) [2018-06-27]. https://numenta. com/resources/biological-and-machine-intelligence/.

[26] 尹相权，曾姗，糜凯．基于人名消歧的自引统计研究 [J]. 情报探索，2015 (5): 57－59, 67. (Yin Xiangquan, Zeng Shan, Mi Kai. Personal name disambiguation-based research on self-citation statistics [J]. Information Research, 2015 (5): 57-59, 67.)

[27] Cortical. io. retina-sdk. py: A python client for the cortical. io retina API [EB/OL]. (2017-03-06) [2018-06-27]. https://github. com/cortical-io/retina-sdk. py.